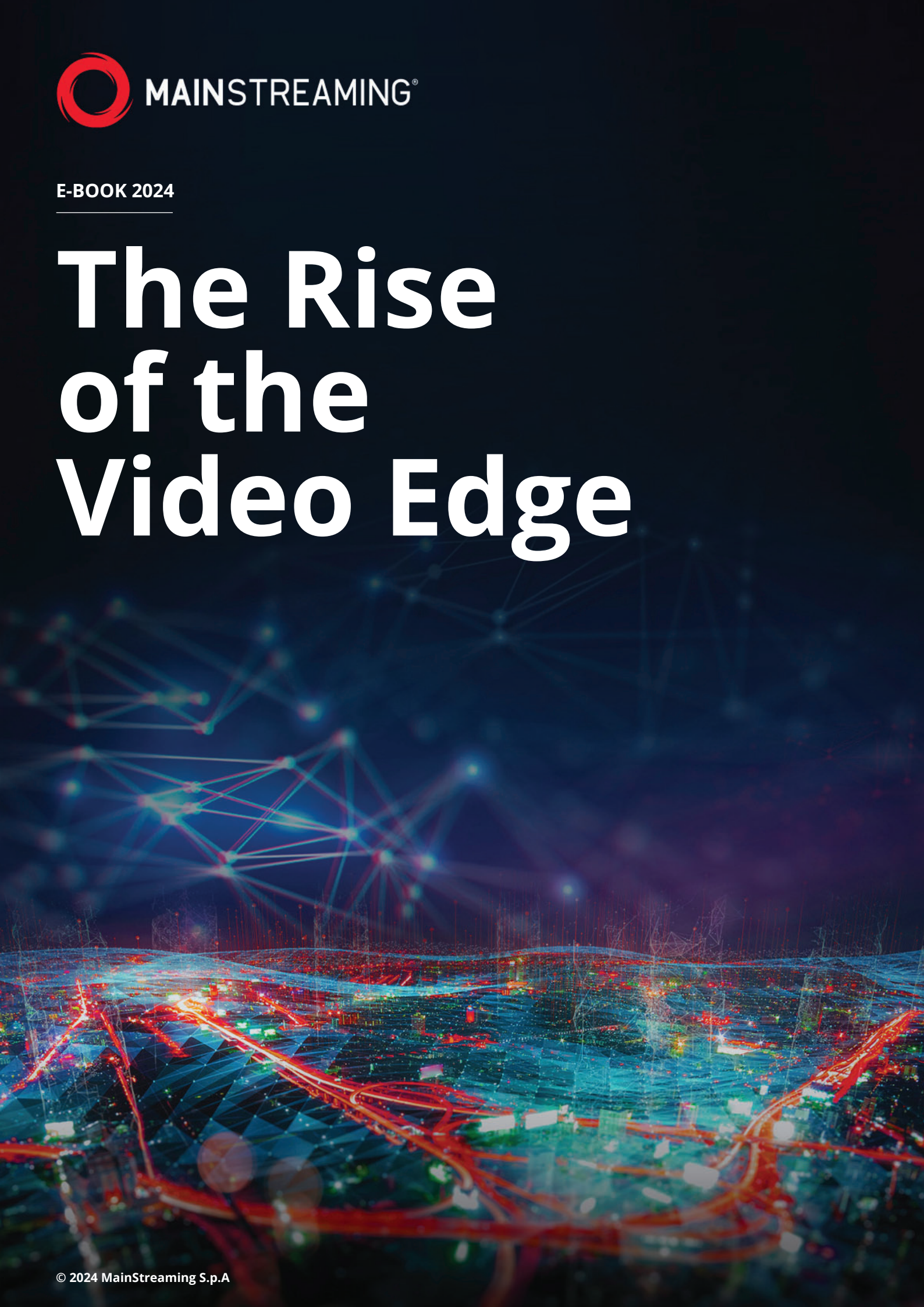
MAINSTREAMING®

E-BOOK 2024

The Rise of the Video Edge



The Discipline of Edge Computing

Edge computing is a computing paradigm, born in the 1990s, which refers to a range of networks and devices at or near the end-user. Edge computing is about processing data ever-closer to where it is being generated or used, which leads to more immediate results for end-users. Edge computing also reduces traffic on centralized core networks and creates greater system resilience from distributed processing operations. Content Delivery was an original use case for edge computing as internet usage increased, to process web files nearer to the end users.

In contrast with Edge computing is Cloud computing. Cloud computing evolved in the 1960s based on network-connected computers and the beginning of the internet. Cloud as we know it today emerged in the mid-2000s as companies like Google promoted the concept of storing and processing data in off-premises rented server capacity. This concept replaced the idea that companies had to store all their data and software on their own hard drives and servers. Cloud computing has been the catalyst for the evolution of rich media services we have today.

As Cloud computing has matured, distributed computing has become standard practice. Resources that are orchestrated in central cloud environments now exist in a bundle of functions that run across multiple servers and computing instances.

Cloud computing is now moving to the Edge.This is a significant step for video streaming services.

For many years, video streaming either involved delivering large static files as downloads to devices or streaming video from servers which breaks down the file into small segments and delivers them to the device for immediate playback. Today, converting broadcasting over to streaming, with all the embedded functionality like closed captions, loudness control, and audio synchronization controls, has pushed the current concepts of cloud computing and edge computing to its limits. Making this even more challenging is the concept of personalized live streaming, where individual live real-time streams can be unique to each viewer.

The limits are stretched because Cloud computing is not time-sensitive. **Edge computing, being closer to the end-user, is more equipped to be time-sensitive.** The Edge is where time matters, and where immediate mass customization can occur. Edge computing eliminates the round trip to a central Cloud platform, reducing system-wide latency. Edge computing keeps the heaviest data processing closer to the end-user to dramatically reduce latency, and leads to automated, immediate decision-making which improves the user experience.

As well as latency benefits, Edge computing is growing because it provides greater resilience to applications by having fewer single points of failure and more points from which service can be maintained. Distributed storage and processing of data enables smaller compute locations to be utilized. More distributed locations enable companies to store and process data in locations that keep them compliant with data and privacy laws.

Content Delivery Networks (CDNs) began as file caching systems to deliver websites and static content to end-users more quickly. This evolved into delivering audio and video files. Today, CDNs process media while delivering it to end-users. **CDNs verify tokens, manipulate manifests, repackage files, apply watermarking processes, and more**. These functions all use compute. This is edge computing, not file caching. This is media service delivery, not just video file delivery.

In September 2023, Forrester published a report entitled "Product Security At The Four Edges" which included a diagram as shown in Figure 1. The Provider edge domain is a base platform on which Telcos provide physical connectivity and hosting space for all other forms of Edge Computing. Enterprise, Operations and Engagement edge domains progressively fan out from large enterprises towards consumers. Video streaming is part of the Engagement edge.

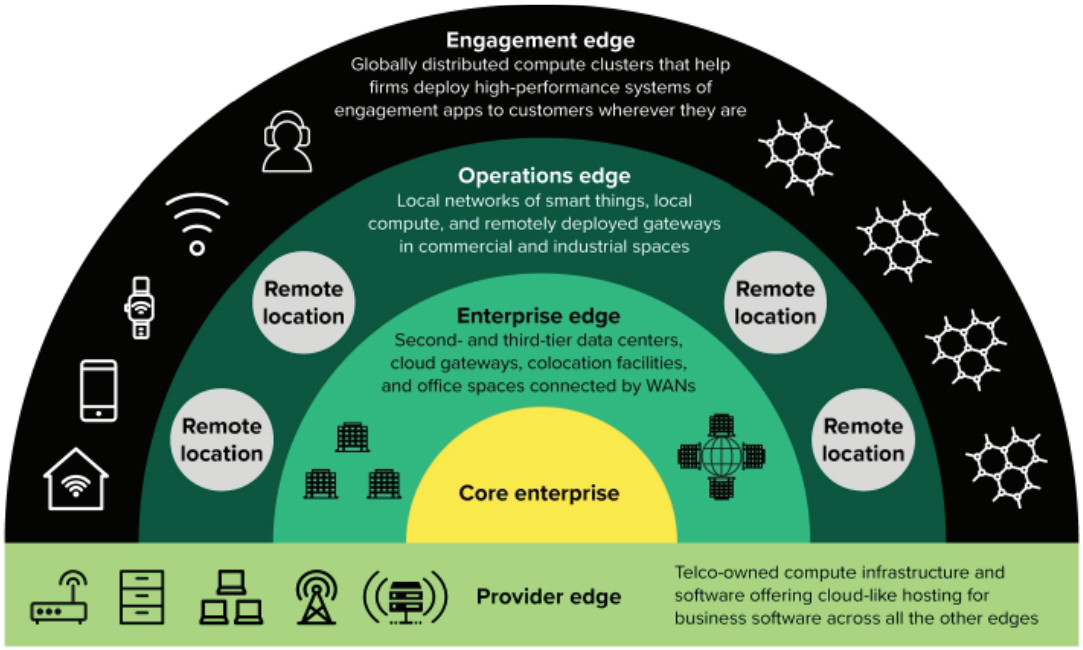


Figure 1: Forrester's Edge Computing domains (Forrester's report "Product Security At The Four Edges", September 2023)

Edge adoption is accelerating. Cloud-native technologies are already well deployed at both Enterprise and Operations edges because enterprises need consistent IT environments to remain efficient in developing and deploying applications. The edge is simply viewed as another place to run application workloads.

Edge adoption is **underpinned by distributed points of presence (PoPs),** but the **innovation today is found in intelligent load analysis across all Edge environments and sophisticated heuristics at the service layer** to determine where to run a workload for maximum performance. This capability is leveraging the innate distributed nature of Edge Computing versus Cloud Computing.

In the Media industry, Edge Computing is leading to more efficient and personalized delivery of video, at scale. It is a key enabler for transferring large TV audiences from satellite, cable, and terrestrial TV networks to IP-based streaming networks.

How Media is Moving to the Edge

Video streaming began as file downloads from a central video storage location to be played from a local device-based file storage system. This was called Video On Demand, or VOD. As consumption grew, the delivery infrastructure evolved to create more local caches that could store files closer to consumers and stream them rather than download them. This minimized the total network bandwidth required as well as the time to wait for video files to be delivered and start playing. But video "streaming" was still about delivering files.

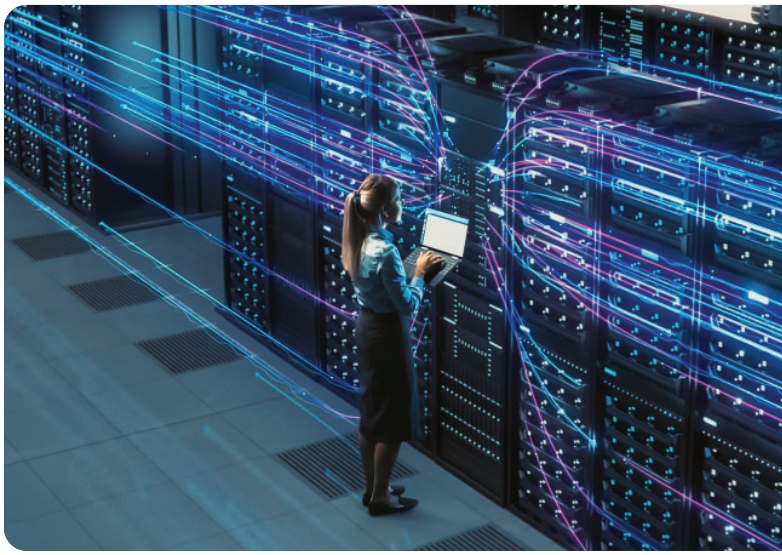
Then live video streaming was added. The network infrastructure could technically support it, even if it had not been designed for it. To achieve a satisfactory viewer experience, latency safety measures had to be introduced to overcome quality issues caused by poor or slow network throughput of the video segments - buffer sizes on video players and video file segment sizes were optimized, and adaptive-bitrate (ABR) technology was invented to help latency-sensitive live video to be correctly delivered over non-deterministic IP networks. As live streaming grew, technology matured to ensure large audiences could be served – architectures for streaming and monitoring became more scalable, specialized software was introduced for controlling streaming sessions, and real-time stream redirection was introduced.

Then rich media services emerged. Video became just one ingredient of the viewer experience. Comments, group chats, voting, betting, and real-time statistics started to enrich VOD content and live event experiences. Video delivery was accelerated by local caches, but the rest of the media service components relied on a centralized or device-level set of service endpoints. Developers focused on scaling the resources in the central Cloud computing location, with some geographical replication to help reduce latency. The advances in media services to address requirements like content discovery, personalization, and multi-device mobility increased the complexity of media services delivery: now each client gets a potentially different set of data, which removes the ability to leverage the same copy of data held in local caches, thereby putting more pressure on live media service delivery which needs to be alleviated with intelligent media service delivery solutions.

Today, the video files we deliver are generally already fully processed into final video assets in centralized locations before being originated into the delivery networks. The processing functions include encoding/transcoding, packaging, and digital rights management. In live video these processes are performed in real-time as the video segments are pushed from studios to playout facilities and towards the audience. In VOD some processes are done ahead of time, like transcoding, and some are often done on-demand and just-in-time, like packaging.

We are now at a point where many media delivery functions can be performed in Edge computing domains. We know how to distribute software and video processing to Edge environments quickly and automatically. More functions are being added at the Edge to make media delivery more efficient and resilient.

But Edge computing in Media is not just about media delivery. It is about delivering the whole Media Service. Therefore, we are also in the process of enabling Media Platform Access Services from the Edge (e.g., user authentication). This requires distribution of the software that we use in the service endpoints: the APIs. This helps organizations delivering media-centric experiences to provide a better user experience and to more effectively protect revenues.



The adoption of Cloud and the Big New Migration to the Edge

Cloud adoption by Enterprise IT and large Media businesses has been progressing for over a decade. In today's Media industry many content processes are running on cloud infrastructure and cloud services, including production, post-production, compliance, acquisition, business-to-business (B2B) content distribution, and direct-to-consumer (D2C) streaming. But we are now seeing the start of the big migration of many of these processes into Edge computing domains. The Edge began as a data delivery environment but is becoming the new Cloud, adding computing to the Edge's already-existing delivery capabilities.

The move to the Edge in the Media industry involves a migration of software and services that drive the Media Services we use. In the never-ending quest to make services faster, better, and cheaper there is a natural assessment of how to leverage the growing capacity of Edge computing platforms, that are expanding in Media primarily to support more streaming video delivery.

The Cloud gave us a new "elastic infrastructure" that replaced fixed server infrastructure. The Edge instead gives us a "hyper-distributed elastic infrastructure" that is even closer to the consumer and can be used for making workloads more efficient and cost-effective.

To use this infrastructure requires a step forward in software development practices that leverage this hyper-distributed serverless state.

In the Media industry, leveraging new Edge architectures is supported by new business models. These use compute capacity at the Edge during off-peak times and bundle media processing into Media Delivery deals. To elaborate, the increasing demand for Media Delivery requires the expansion of streaming delivery system capacity, which includes more memory, compute, and storage resources. At the same time, compute consumption for VOD file preparation remains relatively consistent according to the number of unique videos produced, published, and delivered.

The over-supply of compute capacity due to streaming capacity expansion growth creates an opportunity to better utilize the distributed compute capacity at the Edge for the delivery of Media Services.

The Rise of the Video Edge

Video is already the largest consumer of internet bandwidth. This is expected to continue as today's Broadcasters transition their large audiences to streaming services, and as many retail brands become D2C streamers for communicating to audiences with video. We also expect content formats to evolve from standard 2D screen experiences to multi-screen, immersive, and interactive viewing experiences. This evolution will dramatically increase the amount of content to process and deliver, and the amount of bandwidth required between Edge Platforms and viewers. Video is expected to become an even larger consumer of internet bandwidth. This is likely to require a specialized Video focus to protect viewers' quality of experience and optimize network bandwidth utilization.

The Video Edge has three characteristics that make it uniquely placed to assure service quality, protect revenues, and improve efficiency throughout the Media Services value chain:

1. It is the final secure part of the Media Services chain where media can be controlled and changed.

2. It is the only part of the entire delivery chain that simultaneously manages source content from an Origin and an individual stream for each consumer.

3. It is the only secure part of the delivery chain that is communicating one-to-one with the consumer.

At MainStreaming we are very focused on video. We help our customers to build and deploy Media Services that delight their customers, and that reduce the complexity for developers to build and deploy Rich Media Service Functions at the Edge.

Rich Media Service Architecture & Taxonomy

The Video Edge is about delivering Rich Media Services to consumers. There are multiple technical domains for a video-centric Rich Media Service.

We define the layers in this model as follows:

Media Resource is about audio and video streams and the elements to decode and render them. It includes packaging of video streams, DRM-related operations, and access control and watermarking to fight piracy.

Media Application includes the media controls, for example play, pause, fast forward, seek, 360-degree video panning, and camera switching.

Media Function includes advanced interactions with the media including live/DVR switching, ad insertion, visual seek, graphic overlays, picture-in-picture, and multi-video layouts.

Media Service encompasses all those activities that directly relate to the content such as statistics and community interaction.

Rich Media Service takes the core Media offering and adds significant supplementary offerings, including gaming, betting, and e-commerce. The organizations already providing rich media services leverage their core Media offering to create much larger "universes" related to other products and services they offer. For example, Amazon Prime links viewers into Amazon's shopping environment, while DAZN enables its viewers to buy tickets for sporting events and bet on the games.

Figure 2: Rich Media Service Architecture Layers

How Media Services use the Video Edge

To understand the business benefits the Video Edge can bring to Media Services, we need to describe how the Video Edge is being used today, and how we expect it to be used in the future.

The following diagram shows the core Functions normally used today in Media Services, categorizes them into **Media Service Access Platform functions** vs. **Media Delivery Platform** functions, and highlights the technical domain in which they are typically deployed (i.e., On-Prem, Cloud, Edge, Device).

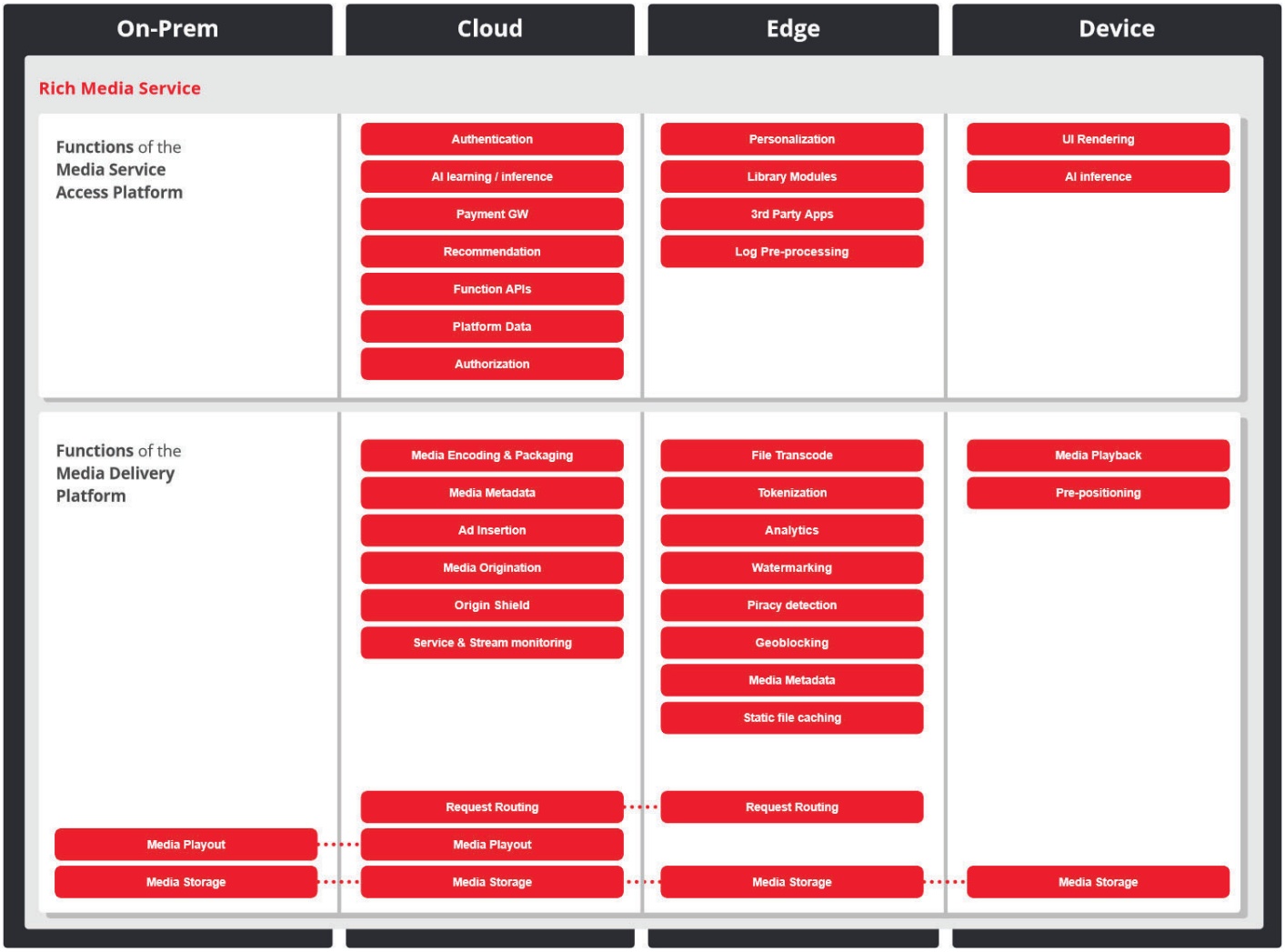


Figure 3: Typical deployment of Media Services functions across four technical domains

As Figure 3 shows, the Media Services Access Platform functions are oriented towards the Cloud, while Media Delivery Platform functions are already being distributed to the Edge. This is logical given the quality, latency, and network bandwidth requirements related to media delivery that exceed those required for media service access. This current orientation is also driven by the centralized databases and many integrated micro-services that work together in a Media Services Access Platform, and the rise of the large Cloud providers that have supported these storage and computing requirements. However, the growth of streaming video and an ever-increasing richness of Media Services means the migration from Cloud to Edge to improve the performance of Media Services is inevitable.

We define the layers in this model as follows:

The **Media Service Access Platform** contains management functions such as authentication and authorization that link to central user databases. AI (Artificial Intelligence) functions operate between central Clouds and Devices, while UI (User Interface) Rendering runs on Devices. This Platform also includes library modules and 3rd party applications that create the Media Services experience on whichever device is being used.

The **Media Delivery Platform** includes the functions that prepare the media to reach devices and play back the media. Encoding, packaging, and inserting advertising all manipulate the media stream. Some of these functions operate in real-time, based on the consumer request, while others, like file transcode, are batch-processed in advance. This Platform also includes the functions that enable video to pass over the internet connection to the viewer, focusing on playout, origination, and monitoring in the centralized On-Prem and Cloud locations, while implementing request routing and security features like watermarking and geoblocking at the Edge, and video playback and pre-positioning of content at the Device. Storage is a function that exists in every part of the Media Delivery value chain.

Functions Moving to the Edge

Some Media Functions have already moved or begun their move to the Video Edge domain. Figure 4 below shows four specific functions. MainStreaming has designed and operated three of these functions at the Edge since the company was originally founded in 2016, with the goal of creating innovative solutions for media delivery when compared to the traditional CDN approach. A fourth function has been added to the Edge more recently based on customer demand and a general approach to reduce ISP core network bandwidth utilization. These functions are how MainStreaming excels in VOD and live video streaming at scale, and why the MainStreaming Video Edge Platform is ready to support more and more Edge functions for Rich Media Services.

1. **Service and Streaming Monitoring**
   By running the Service and Streaming Monitoring function at the Edge, monitoring capacity expands with Edge capacity expansions. It is not a separate expansion that constrains total system capacity and adds additional cost. By monitoring on every Edge, the total system is more responsive to changes in conditions at each Edge, and therefore adapts very well for the most sensitive streaming use case of live video streaming.

2. **Origin Shield**
   Distributing the Origin Shield function to run on the Edge means less hardware needs to be deployed and less content needs to be stored in the overall delivery system. This results in more efficient infrastructure, including lower energy consumption and lower CO2 emissions. To ensure running the Shield function on the Edge protects the Origin from being overloaded with requests, this function includes the ability for any Edge to be the Master Edge for communicating with the Origin for individual channels or pieces of content. MainStreaming's patented technology for ultra-low latency intra-server communications means that the network of Edges can quickly communicate together and supply content to each other.

3. **Function APIs**
   MainStreaming has over 300 Function APIs used to manage its Video Edge network, which can also be exposed to higher-level systems. These APIs were originally designed to run in the Edge domain. They are mostly used by MainStreaming but will be progressively exposed to customers to support plans to run edge computing workloads directly on the MainStreaming Video Edge network.

4. **Repackaging**
   The Repackaging function has been introduced based on customer demand. Delivery from the Origin with pre-packaged content in HLS format has evolved to only deliver a mezzanine CMAF format from Encoder to Origin to Edge, which is then repackaged from CMAF to HLS for final device delivery at the Edge. This processing step at the Edge introduces three efficiency and reliability improvements. First, it offloads processing from the Origin environment, which protects against Origin overload and improves Origin reliability. Second, it reduces the number of delivery variants that need to be passed from Origin to Edge, which reduces bandwidth consumption between Origin and Edge. And third, it avoids unnecessary processing of all forms of package type when perhaps not all forms will be requested.

These 4 functions run at the Edge today in MainStreaming's Video Edge network, specifically to improve the performance of media delivery.

Figure 4: The Media Service functions already moved or currently moving into the Edge Computing domain.

The next four functions that are currently moving to the Video Edge domain are being driven by customer demand in a continuous drive to improve Media Services. These four functions are based on current MainStreaming visibility of customer priorities, but new priorities continue to emerge from customers through collaborative feedback.

Cloud-based functions currently tie into the centralized computing environment of Cloud platforms, often for historical reasons as the large public Cloud providers have been the first to commercially offer this sort of platform. Device-side functions currently use the on-board computing power of devices – which are becoming more powerful – to perform key functions that are too personalized or too low-latency to run in Cloud domains. The general downside with Device-side functions is that they consume power on every device in which they run. The rise of the Video Edge has created a superior alternative for running Media Service Access Platform functions.

1. **Recommendations (from Cloud to Edge)**
   As personalization of Media Services increases and the size of audiences on streaming platforms grow, the speed of serving recommendations to consumers needs to keep up. Utilizing the lower-latency location of the Video Edge allows algorithms on a per-user basis to run and respond more quickly to requests and distributes the Recommendation Function which reduces the risk of an outage affecting all Platform users.

2. **User Interface (UI) Rendering (from Device to Edge)**
   UI Rendering is typically performed on the viewing device. This compositing process generally creates a heavy load on the device's CPU. But for legacy devices that are not powerful enough to run all the UI's features then UI rendering could be offloaded to the Edge to pre-render and send to the client as images. This legacy device support situation is mostly experienced by public broadcasters that have requirements to provide their streaming services to a very large majority of the population.

3. **Platform Data to Data State (from Cloud to Edge)**
   Performance of the Media Service Access Platform is critical to impress and retain viewers. Therefore, Platform Data is moving from the central Cloud into a Data State function on the Video Edge to improve performance. This enables all Edges to be fully updated within milliseconds with the latest data-state held centrally in the Cloud. By doing this, an Edge can run a business logic function (i.e., an API) and use the Data State to check the latest data. This helps make business decisions faster compared to returning to centralized Cloud locations.

4. **Artificial Intelligence (AI) Inference Offload (from Cloud & Device to Edge)**
   AI Inference is an area in which Apple and Google are leading the way. Mostly this function runs in the Cloud with a small percentage running on the device, based on inputs from a centralized Cloud-based AI learning environment. But like UI Rendering, if the device cannot manage the intensive computing, then the AI Inference can instead run on the Edge and be delivered to the device. And like Ad Insertion, if the latency in the Cloud is too slow for the user experience, then AI Inference can be run on the Edge, closer to the device.

5. **Advertizing (Ad) Insertion (from Cloud to Edge)**
   Ad Insertion that is managed centrally and applied to all viewers receiving the stream is moving to more personalized Ad Insertion that is managed at the Edge. Through manifest manipulation, Video Edges can request and insert unique Ads specifically for individual viewers. Cloud based Ad Insertion has been observed to be slow, while Client-side Ad Insertion has low market value because of device-level ad blockers that advertisers want to avoid. Network-side Ad Insertion is therefore seen as the superior solution, which for latency reasons should run at the Edge.

Other Cloud-based and Device-based functions are already being considered or planned for migration to the Video Edge domain. For example, Authentication and Authorization of users onto the Media Service would make sense to manage in a distributed manner at the Edge. One of the main challenges with large live event streaming is to authorize many thousands or millions of users within a few minutes of the live event starting. Distributing this workload could potentially resolve this challenge. While a potential candidate to move from the Device domain to the Edge is channel changing within the UI.

Leading Streaming Providers are looking closely at the potential benefits of Edge Computing, for both Media Delivery and Media Service Access. As we observe the growth of streaming delivery capacity in Edge Networks to support larger streaming audiences, we are simultaneously observing a growth of computing capacity in these Edge Networks. This computing capacity can be, and should be, leveraged to provide the new efficiencies and greater resilience required by today's Media Services.

Glossary - main functions

Authentication: A security process that verifies the identity of a user or system, ensuring that only authorized individuals can access certain media services or information.

Authorization: The process of granting or denying specific permissions to authenticated users, determining what actions they are allowed to perform within a media service platform.

Payment Gateway (GW): A system that interacts with external system in secure way to accept debit or credit card, and additional payment methods accepted by purchases from customers, facilitating the secure transfer of payment information from the media service to the financial networks.

AI Learning Interface: An interface designed to interact with artificial intelligence systems, enabling the improvement of media services through machine learning by feeding data and interpreting results.

Log Preprocessing: The initial step in log analysis involving the cleaning, structuring, and organizing of information and media access log data to make it suitable for further analysis or processing in media service operations.

Personalization: The customization of media content offer and services to fit the individual preferences and behavior of users, enhancing user experience by delivering content that is most relevant to them.

Recommendation: A software system within media services that suggests content to users based on their preferences, viewing history, and other user-specific data, aiming to enhance engagement and satisfaction.

Library Modules: Reusable code components within a media service platform, providing specific functionalities, which can be easily integrated into different parts of the system.

3rd Party Apps: External applications integrated into a media service platform, offering additional features or content that are not developed by the platform's original provider.

UI Rendering: The process of converting UI code into interactive graphical interface across various devices such as Smart TVs, set-top boxes, game consoles and mobile phones. This involves displaying graphical elements like buttons, menus as well all additional information like images and metadata related to audiovisual content.

Media Encoding: The process of converting audiovisual content into a digital format using specific codecs. Encoding usually altering media size and format to make the media suitable for streaming, storage, or transmission, while trying to maintain balance between quality and size.

Media Packaging: The process of wrapping encoded media content into format suitable to storage, transmission or streaming. Packaging may include segmenting the content for adaptive streaming and implementing encryption of content for DRM (Digital Rights Management) for content protection. This enables the delivery of media content over the network in formats compatible with various playback devices and streaming protocols, such as HLS (HTTP Live Streaming) and DASH (Dynamic Adaptive Streaming over HTTP).

Ad Insertion: The process of integrating advertisements into content. Ad insertion can be done in real-time (dynamic ad insertion) or pre-embedded into the content (static ad insertion). Dynamic ad insertion allows for advertising to be served to different users based on their preferences, demographics, viewing behavior and/or geographic location, enhancing the relevance of ads.

Service and Stream monitoring: The continuous oversight of a video platform's performance, stability and availability, focusing on metrics critical to delivering a high-quality viewing experience. This may include monitoring streaming quality, video load times, server response times, bandwidth usage, etc. The goal is to proactively detect and address issues such as buffering, service downtime, or quality of service degradation.

Origin Shield: Origin Shield is a caching layer placed between a content delivery network (CDN) and the origin server that hosts the original content. This intermediary layer serves as a shield to reduce the number of direct or indirect (via CDN) requests to the origin.

Request Routing: Request routing within a Rich Media Service involves dynamically directing user requests to the most suitable edge server in order to deliver content as quickly and efficiently as possible. The criteria for selecting an edge server can include geographic proximity to the user, server load, network congestion. The Request Routing ensures that users receive data from the nearest or most optimal location, significantly reducing latency and improving the overall speed and reliability of content delivery.

Transcoding: The process of converting audio or video files from one format or codec to another. This is often necessary to ensure compatibility across various playback devices, platforms, and network conditions. Transcoding can involve changing the file format, compressing or decompressing the content, altering the bit rate, or resizing the dimensions of video content.

Tokenization: is a security measure to control access to content. The process involves generating a unique token, for example string of characters, that is validated by the CDN or Platform before allowing access to the requested content. The token is often included in the URL or request header and may carry information such as the expiry time, IP address restrictions, and allowed referrer URLs. Tokenization in the CDN context is particularly useful for media streaming services, online gaming, and other digital content providers who need to secure their content against piracy and ensure that only authorized users can access premium or restricted content.

Watermarking: the technique of embedding a unique, identifiable pattern or “mark” into digital media, such as images, videos, or audio tracks, without significantly altering the original content. This mark can be invisible, embedded within the content’s data in a way that it can be detected only with particular process. Watermarking serves for track distribution and deterring unauthorized sharing and providing a means of identifying the source of pirated content.

Media Metadata: Data that describes media content. It may include details like title, artist, duration and description of content. Metadata facilitates content organization, management, and discovery across digital platforms, making it essential for streaming services and media libraries.

Geoblocking: The process of restricting access to content based on a user’s geographic location. It’s used to enforce licensing agreements, comply with local laws, and manage content distribution.

Piracy Detection (Server Side): A method of identifying unauthorized distribution of copyrighted content by monitoring server activity and data. This approach focuses on detecting piracy directly from the platform hosting the content, analyzing access patterns, download frequencies, and other metrics that may indicate illegal sharing or downloading. Server-side piracy detection enables content providers to proactively safeguard their digital assets by identifying and mitigating piracy incidents at their source, often before they spread across the internet.

Media Playback: The act of playing audio or video files on a device, allowing users to view or listen to multimedia content. It involves decoding digital media for presentation on various devices like computers, smartphones, Smart TV’s and STB’s.

MAINSTREAMING®

We Stream The Future